Confidence Measure by Substring Comparison for Automatic Speech Recognition

Bartosz Ziółko, Tomasz Jadczyk, Dawid Skurzok, Mariusz Ziółko*
Department of Electronics,AGH University of Science and Technology
Al. Mickiewicza 30, 30-059 Kraków, Poland
www.dsp.agh.edu.pl
{*bziolko, jadczyk, skurzok, ziolko*}*@agh.edu.pl*

Abstract

Two possible confidence measures for automatic speech recognition are presented along with results of tests where they were applied. One of them is widely known and it is based on comparing the strongest hypotheses with an average of a few next hypotheses. We found it not efficient in all cases, this is why we came up with our own method based on comparison of substrings. New algorithm was found useful in real applications for spoken dialogue system, in a module asking to repeat a phrase or declaring that it was not recognised. The method was designed for Polish language, which is morphologically rich. The method is tuned to situations in which there are several similar utterances in a dictionary.

1. Introduction

Research on automatic speech recognition (ASR) started several decades ago. Most of the progress in the field was done for English. It has resulted in many successful designs, however, ASR systems are always below the level of human speech recognition capability, even for English. In case of less popular languages, like Polish (with around 60 million speakers), the situation is much worse [1–4]. Polish speech contains high frequency phones (fricatives and plosives) and the language is highly inflected and non-positional.

In a dialogue based application it is crucial not only to provide a hypothesis of what was spoken but also to evaluate how likely it is. A simple probability is not always a good measure because its value depends on too many conditions. In case of dialogue systems, additional measure evaluating if the recognition is creditable or not, is very useful. A relation to other, non-first hypothesis can provide it. It allows to repeat a question by a spoken dialogue system or choose a default answer for an unknown utterance. The purpose of confidence measures is to estimate the quality of a result. In speech recognition, confidence measures are applied in various manners.

Existing types and applications of confidence measures were well summarised [5–7]. Confidence measures can help to decide to keep or reject a hypothesis in keyword spotting applications. They can be also useful in detecting out-of-vocabulary words to not confuse them with some similar vocabulary words. Moreover, for acoustic adaptation, confidence measure can help to select the reliable phonemes, words or even sentences, namely those with a high confidence score. They can be also used for the unsupervised training of acoustic models or to guide a dialogue in anserwing services in order to require a confirmation only for words with a low confidence score. Recently applying Bayes based confidence measure for reinforced learning was also tested [8]. Confidence measures were also applied in a new third-party error detection system [9]. Confidence measures are even more important in speaker recognition. A method based on expected log-likelihood ratio was recently tested in speaker verification [10].

Confidence measures can be classified [7] according to the criteria which they are based on:

- semantic,
- language modelling,
- acoustic stability,
- hypothesis density,
- duration,

*This work was supported by NCBiR grant 0021/R/D2/201/01 O ROB 0021 01/ID 21/2

- likelihood ratio,
- lattice-based posterior probability.

2. Literature Review

Let us follow with summarising some results and views on confidence measures for speech recognition which we have found in the latest papers. In some scenarios it is very important to compute confidence measures without waiting for the end of the audio stream [7]. The frame-synchronous ones can be computed as soon as a frame is processed by the recognition system and are based on a likelihood ratio. They rely on the same computation pattern: a likelihood ratio between the word for which we want to evaluate the confidence and the competing words found within the word graph. A relaxation rate to have a more flexible selection of competing words was introduced.

Introducing a relaxation rate to select competing words implies managing multiple occurrences of the same word with close beginning and ending times. The situation can be solved in two ways. A summation method adds up the likelihood of every occurence of the current word and adds up the likelihood of every occurence of the competing words. A maximisation method keeps only the occurence with the maximal acoustic score.

The frame-synchronous measures were implemented in three ways regarding a context: unigram, bigram and trigram. The trigram one gave the best results on a test corpus.

The local measures estimate a local posterior probability in the vicinity of the word to analyse. They can use data slightly posterior to the current word. However, this data is limited to the local neighbourhood of this word and the confidence estimation does not need the recognition of the whole sentence. Local measures gave better results on a test set.

Two n-gram confidence measures based evaluations were also recently tested [11]: 7-gram based on part-of-speech (POS) tags and 4-gram based on words. The latter was not succesfull in detecting wrong recognitions. Applying POS tags in a confidence measure was succesful, possibly because it enables analysis on larger time scale (7-gram instead of 4-gram).

A new phonetic distance based confidence measure was suggested [12]. It applies distances between subword units and density comparison (called anti-model by authors). The method employs separate phonetic similiarity knowledge for vowels and consonants, resulting in more reliable performance. Phonetic similarities between a particular subword model and the remaining models are identified using training data

$$P(X^{\{i\}}|\lambda_{i,1}) \geq P(X^{\{i\}}|\lambda_{i,2}) \geq \cdots \geq P(X^{\{i\}}|\lambda_{i,M}) \quad (1)$$

where $X^{\{i\}}$ is a collection of training data labeled as model λ_i and $\lambda_{i,m}$ indicates the mth similar model among M subword models compared to the pivotal model λ_i.

3. 1-to-3 Comparison

The most widely known confidence measure is of hypothesis density type. It compares the strongest hypothesis with an average of the following n weaker ones by dividing (Fig. 1). In our experiments $n = 3$ was empirically found useful and it is a common value for this parameter in other systems as well. Our evaluations were done for sentence error rate. In the first evaluations it worked very well but later on, we found out, that its usefulness is limited in real dialogue applications because it had similar ratio for sentences allowed by a dictionary as for the ones which were not allowed. It was confirmed in later statistical tests with larger dictionaries.

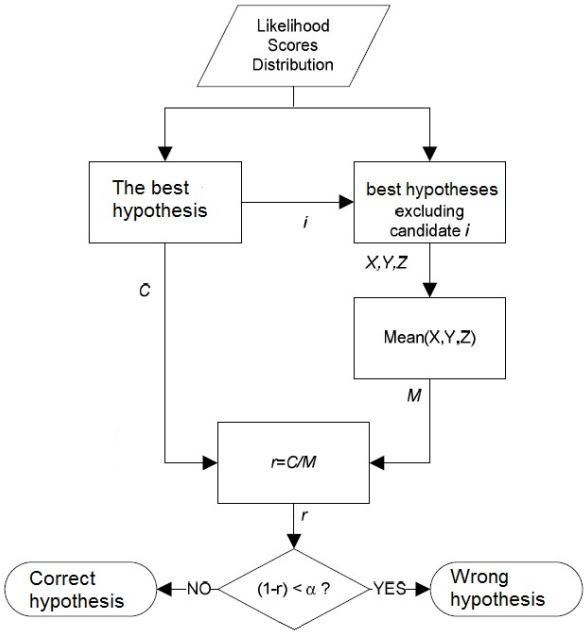

Figure 1. Algorithm of a standard method of confidence measure

4. Substring Comparison

Our new confidence measure was designed and implemented for real application scenarios where there are several utterances very similar to each other. Such situation is especially common in morphologicaly rich languages like Polish [4], Czech [13] or Finnish [14]. In this type of scenarios classical confidence measures frequently fail to help detect wrong recognitions. Our new approach operates by comparing substrings of phonemes of the strongest hypothesis with the following ones. The method was implemented for a spoken dialogue system and it evaluates whole sentences in a dialogue so time alignements are not considered.

First, we make sure that the following hypothesis has a higher cost and different phonetic transcription to avoid numerical problems in the algorithm later on. In some rare cases two hypotheses can have the same cost by a coincidence. In other cases it is possible that orthographicaly different words have the same phonetic transcriptions and this is why two hypotheses with the same transcriptions can appear.

In the next step the strongest, but not first, hypothesis n which is not phonetically similar to the first one has to be found. Two phonetic transcriptions are similar if any of them is an exact substring of the other. Then a difference

$$d = p_1 - p_n \tag{2}$$

between probabilities of the primary hypothesis and the nth - found in the previous step has to be calculated. So for example, if the strongest hypothesis is /abc/, the second one is /abcd/, and the third one is /xyz/, the probabilities of the first and the third will be compared. A real case example is presented on fig. 2.

Then the confidence score is calculated

$$c(d) = \begin{cases} 0.1 & \text{for} \quad d \leq 0.001 \\ 0.1 + 50 * d & \text{for} \quad 0.001 < d \leq 0.01 \\ 0.6 + 25 * (d - 0.01) & \text{for} \quad 0.01 < d \leq 0.026 \\ 1 & \text{for} \quad d > 0.026 \end{cases}, \tag{3}$$

where all numerical values were found empircally during the development experiments.

As it can be concluded, the suggested substring method is quite a new approach, which does not fall directly into any of the confidence measure types presented above and listed in literature [7].

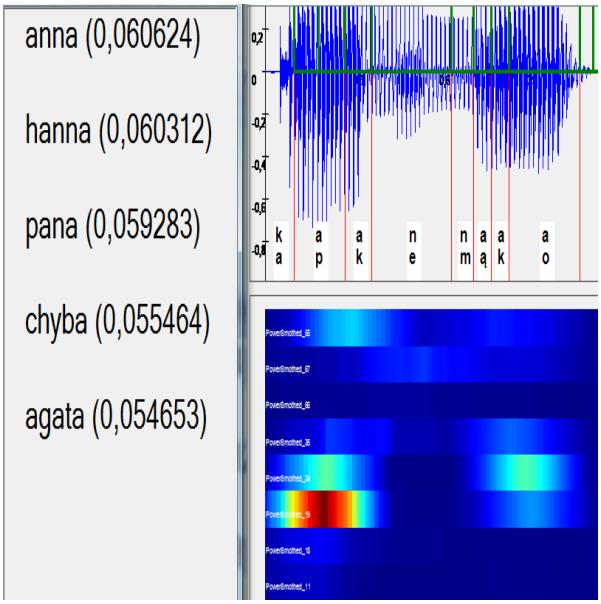

Figure 2. A screenshot from the developers version of our ASR system presents an example of how the decribed phonetic substring confidence measure can be applied. The left part shows the ranking of top 5 hypotheses and the right one, the time and frequency representation of the analysed audio file. If the probabilities of the strongest hypothesis "Anna" and the following one "Hanna" were substracted, the difference would be $d = 0.000\ 312$, giving final score $c = 0.1$, which is very low. But "Anna" is a substring of "Hanna" (both are female names). This is why instead of the second hypothesis, the third one "pana" can be used to compare with. It results in $d = 0.001\ 341$ and a final score $c = 0.17$, which is stil quite low. However, the window shows ortographic transcriptions, and "anna" transcription can be /ana/ which is a substring "pana" transcription /pana/. Then, let us compare "Anna" with "chyba". It gives $d = 0.005\ 16$ and the final substring comparison confidence measure $c = 0.36$

5. Tests and Results

The standard 1-to-3 method was compared with the substring method in a sequence of experiments on our

own test corpus. The recordings consists of over 100 audio files, each with one sentence, spoken by the same male speaker. All tests were made using AGH ASR system [4].

The audio files have sampling rate 16 000 Hz and 16 bits per sample. The sentences have mainly political context (disscusions and public speeches) and Kaczmarski songs. They were recorded in a regular, not noisy office from prepared sentences. No language model was used in the tests. Each recording is a phrase or a sentence. The shortest one is 'proszę o ciszę' (Eng. silence please) and the longest one 'Polska doświadczyła takiej katastrofy na początku lat dwudziestych i ponownie w tysiąc dziewięćset osiemdziesiątym dziewiątym roku' (Eng. Poland experienced such a disaster in the beginning of twenties and again in ninteen eighty nine). In average the sentences have around 8 words. The recordings last 6 minutes in total. All error rates are given for whole sentences. The test sentences were not used during development which was based on on-line recordings, directly from a microphone.

In the first experiment, a dictionary with sentences from the recordings was connected to the system. 79.04% of recordings were correctly recognised. An average position of the correct hypothesis in the ranking of all hypotheses was 7.7, which is suprisingly high with the mentioned correctnes of recognition. It is a result of a few recordings which were very badly recognised.

1-to-3 confidence measure method gave score 0.95 in average. The substring method resulted in score 0.88 in average. It has to be mentioned that scores of both methods are uncorelated and difficult to compare directly with each other as they have completely different algorithms (the scale of changes can be different). However, even though, they are not scaled in the same way, both evaluate recognitions more creditable for higher values of the score.

In the second experiment, a different dictionary was used. It was a dictionary of Grocholewski CORPORA [15] sentences. This experiment was conducted to evalute the confidence measure behaviour for wrong recogntions. Of course, it resulted in 0% recognition because none of the spoken sentences was in the dictionary. The more interesting are the obtained confidence scores. 1-to-3 method gave 0.21, and the substring method 0.93.

In the experiment described above, the standard 1-to-3 method clearly showed that recognition were wrong, while the substring method failed giving better score then in the experiment with the correct dictionary. However, the experiment described above is not very realistic for commercial solutions. The dictionary does not contain similar utterances. In a real world scenario, especially in higly inflected languages like Polish, there are huge number of very similar utterances.

This is why in a human tests (made by using the final commercial solution which uses our system with different possible settings), 1-to-3 method was found not succesful. This situation was simmulated by the third scenario, where we combined all possible dictionaries available in our system (around 20 000 utterances) apart from the dictionary with the spoken sentences. The recognition rate was again 0%. In this case, 1-to-3 method gave score 0.98 which is better then for the experiment with correctly recognised recordings.

In contrary, the substring method, gave score 0.46 which is much less then the score for the correct recognition experiment. The same observation can be found in some studies, that likelihood ratio-based measures give satisfactory performance but they were evaluated for small vocabulary ASR. A good summary is given in [7].

6. Conclusions

The suggested confidence measurement method based on substring comparision works much better then the classical 1-to-3 method in an experiment motivated by real applications and end-user tests. The method was designed for morphologically rich languages, as it gives better scores if the strongest hypotheses are phonetically similar. At the current version the applied measure of phonetic similarity is very basic and this is why it is planned to be extended to a more sophisticated one.

References

[1] G. Demenko, S. Grocholewski, K. Klessa, J. Ogórkiewicz, A. Wagner, M. Lange, D. Śledziński, and N. Cylwik, "JURISDIC – Polish speech database for taking dictation of legal texts," *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 1280–1287, 2008.

[2] L. Pawlaczyk and P. Bosky, "Skrybot - a system for automatic speech recognition of Polish language," *Advances in Soft Computing, Man-Machine Interactions, Springer*, vol. 59/2009, pp. 381–387, 2009.

[3] K. Marasek, L. Brocki, D. Koržinek, K. Szklanny, and R. Gubrynowicz, "User-centered design for a voice portal," *Aspects of Natural Language Processing, Lecture Notes in Computer Science 5070*, pp. 273–293, 2009.

[4] M. Ziółko, J. Gałka, B. Ziółko, T. Jadczyk, D. Skurzok, and M. Mąsior, "Automatic speech recog-

nition system dedicated for Polish," *Proceedings of Interspeech, Florence*, 2011.

[5] F. Wessel, R. Schlüter, K. Macherey, and H. Ney, "Confidence measures for large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 288–298, 2001.

[6] G. Guo, Ch. Huang, H. Jiang, and R.-H. Wang, "A comparative study on various confidence measures in large vocabulary speech recognition," *Proceedings of International Symposium on Chinese Spoken Language*, pp. 9–12, 2004.

[7] J. Razik, O. Mella, D. Fohr, and J.P. Haton, "Frame-synchronous and local confidence measures for automatic speech recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 25, pp. 157–182, 2011.

[8] C. Molina, N.B. Yoma, Claudio Garretn F. Huenupán, and Jorge Wuth, "Maximum entropy-based reinforcement learning using a condence measure in speech recognition for telephone speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 5, pp. 1041–1052, 2010.

[9] L. Zhou, Y. Shi, and A. Sears, "Third-party error detection support mechanisms for dictation speech recognition," *Interacting with Computers*, vol. 22, pp. 375–388, 2010.

[10] R. Vogt, S. Sridharan, and M. Mason, "Making confident speaker verification decisions with minimal speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 6, pp. 1182–1192, 2010.

[11] S. Huet, G. Gravier, and P. Sébillot, "Morpho-syntactic post-processing of n-best lists for improved french automatic speech recognition," *Computer Speech and Language*, vol. 24, pp. 663684, 2010.

[12] W. Kim and J.H.L. Hansen, "Phonetic distance based condence measure," *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 121–124, 2010.

[13] J. Nouza, J. Zdánský, P. David, P. Cerva, J. Kolorenc, and D. Nejedlová, "Fully automated system for czech spoken broadcast transcription with very large (300k+) lexicon," *Proceedings of INTERSPEECH*, pp. 1681–1684, 2005.

[14] T. Hirsimaki, J. Pylkkonen, and M. Kurimo, "Importance of high-order n-gram models in morph-based speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17(4), pp. 724–32, 2009.

[15] S. Grocholewski, "First database for spoken Polish," *Proceedings of International Conference on Language Resources and Evaluation, Grenada*, pp. 1059–1062, 1998.